Best practice manual for findability, re-use and accessibility of infrastructures

Deliverable D1.3

6 December 2022

Authors

Wouter Addink[1], Niki Kyriakopoulou[1], Lyubomir Penev[3,4], David Fichtmueller[5] , Ben Norton[6], David Shorthouse[7]

[1] *Naturalis Biodiversity Center, Leiden, Netherlands*
[2] *Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands*
[3] *Pensoft Publishers, Sofia, Bulgaria*
[4] *Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria*
[5] *Botanic Garden and Botanical Museum Berlin Dahlem, Berlin, Germany*
[6] *North Carolina Museum of Natural Sciences: Raleigh, NC, US*
[7] *Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada*

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Best practice manual for findability, re-use and accessibility of infrastructures
Deliverable n°:	D1.3
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP1
Lead beneficiary:	Naturalis Biodiversity Center
Citation:	Addink, W., Kyriakopoulou, N., Penev, L., Fichtmueller, D., Norton, B. & Shorthouse, D. (2022). *Best practice manual for findability, re-use and accessibility of infrastructures*. Deliverable D1.3 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 18
Actual submission date:	6 December 2022

Deliverable status:

Version	Status	Date	Author(s)
1.1	Final	30 Nov 2022	all
1.0	Final/Draft	28 Nov 2022	Wouter Addink, Niki Kyriakopoulou RI forum members and external experts
0.1	Initial Draft, APIs	27 June 2022	Wouter Addink
0.2	API recommendations discussed with technical forum	29 June 2022	Wouter Addink, technical RI forum members

Table of contents

Preface

This manual describes **29 best practices** for findability, re-use and accessibility of biodiversity data hosted by research infrastructures with more than **60 recommendations** for implementation. The best practices have been described as concise as possible and numbered for easy reference, and can be used in combination with the recommendations for interoperability among infrastructures described in deliverable D1.2. The best practices summarise the results of discussions with infrastructures in the project to agree on best practices for findability, re-use and accessibility of their services, with input from external experts. The best practices focus on API services, which play a key role in linking data between the infrastructures and creating a network of knowledge, but also describe more generic best practices on e.g. modalities of access, building communities and trust. The best practices aim to improve the findability, re-use and accessibility for the infrastructures themselves, but also for other users of the services: Researchers, Developers, Citizen scientists and Data providers. For most best practices one or more recommendations are given for their implementation.

The best practices will be made available through the Biodiversity Knowledge Hub (BKH), also developed through BiCiKL. For this purpose they have already been grouped per user group in appendix I to fit the future 'information and guidelines' pages of the BKH targeted to different users. Appendix II gives an overview of infrastructure services and their modalities of access, and appendix III provides an initial inventory of the API services and their compliance with the best practices to guide further improvement of the services.

Summary

United and coordinated efforts of biodiversity data infrastructures are needed to bring together various data forms from many different scientific areas. Biodiversity data are considered of great importance and use when they form a network of knowledge that can be seamlessly integrated and presented to various audiences, promoting both research and education. The Biodiversity Community Integrated Knowledge Library (BiCIKL) project seeks to maximise the potential of integrated data sources by striving to connect fragmented data derived from biological, paleontological, and geological specimens and collections, as well as all derived information such as literature in the form of taxonomic treatments, research papers etc., taxonomic information and molecular sequences provided by these infrastructures, under the umbrella of common digital practices and policies in curation, data sharing and open data access over different scientific fields.

One of the main goals of BiCIKL is to create bi-directional links between various data types, a process enabled by: a) the adoption of globally unique and persistent identifiers upon agreement among all stakeholders, that link to digital specimen objects, collections, taxonomic treatments, people, sequence data and taxa, and b) implementation of the best practices for the generation, management and curation of interlinked data by the host infrastructures. At the same time, infrastructures should be readily discoverable and accessible by end users, providing data that enable re-usability. In this manual we give an overview of the best practices and their associated recommendations for infrastructures on making the most out of their services and data, for establishing a network of knowledge with other infrastructures, for servicing researchers, data providers and other end users. These

guidelines have been developed in collaboration with the infrastructures through Technical RI Forum meetings organised in the context of the BiCIKL project.
Practices and recommendations were divided into six categories: 1) modalities of access, 2) building communities and trust, 3) technology and standards, 4) versioning of APIs and their data, 5) bi-directional linking between infrastructures and 6) API design patterns and naming conventions. A second division into three user groups (Infrastructures, Data providers, Users e.g. Researchers, Developers and Citizen scientists) is presented in Appendix I.

List of abbreviations

ABCD	Access to Biological Collection Data, a standard for access to and exchange of data about specimens and observations
AGU	American Geophysical Union
API	Application Programming Interface, a software interface for two or more computer programs to communicate with each other
BiCIKL	Biodiversity Community Integrated Knowledge Library, a project funded by the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492
BKH	Biodiversity Knowledge Hub, a one-stop access point to guidelines, standards, data and services from 15 research infrastructures, under development in the BiCIKL project.
CC-BY-NC	Creative Commons Attribution-NonCommercial, a license for sharing material
CC-Zero	Creative Commons Zero, a 'no rights reserved' public domain mark.
CETAF	Consortium of European Taxonomic Facilities
CSV	Comma-Separated Values, a delimited text file that uses a comma to separate values
DOI	Digital Object Identifier, a persistent identifier or handle used to uniquely identify various objects, standardized by the International Organization for Standardization (ISO)
DOIP(v2)	Digital Object Interface Protocol, a protocol that specifies a standard way for clients to interact with digital objects
DwC	Darwin Core, a standard to facilitate the sharing of information about biological diversity
EOSC	European Open Science Cloud, Europe's vision to deliver a web of FAIR data and related services for research
EU	European Union
FAIR	Four foundational principles to improve Findability, Accessibility, Interoperability, Reusability of digital assets as guide to data producers
GBIF	Global Biodiversity Information Facility

GUID	Globally Unique Identifiers (also known as Universally Unique Identifiers', or UUIDs) are 128 bit integers represented as 36-character randomised strings that follow the RFC 4122 specification.
HTML	HyperText Markup Language, the standard markup language for documents designed to be displayed in a web browser
HTTP	Hypertext Transfer Protocol, a set of rules for transferring files over the web
HTTPS	Hypertext Transfer Protocol Secure, an extension of the Hypertext Transfer Protocol (HTTP) used for secure communication over a computer network
IG	Interest Group
ISO	International Organization for Standardization
JSON	JavaScript Object Notation, a lightweight data-interchange format, easy for humans to read and write and easy for machines to parse and generate
JSON-LD	JavaScript Object Notation for Linked Data, a JSON schema specifically created to facilitate the exchange of linked data.
MIME type	A media type (Multipurpose Internet Mail Extensions) that indicates the nature and format of a document, file, or assortment of bytes
OAuth2	industry-standard protocol for authorization for APIs
OGC	Open Geospatial Consortium
PID	Persistent Identifier, a long-lasting reference to a document, file, web page, or other object that is globally unique, persistent and resolvable.
REST	Representational State Transfer, a software architectural style to describe a machine-to-machine interface
RI	Research Infrastructure
TDWG	Biodiversity Information Standards (Taxonomic Databases Working Group)
URI	Uniform Resource Identifier, a unique sequence of characters that identifies a logical or physical resource used by web technologies.
URL	Uniform Resource Locator, a web address, a reference to a web resource that specifies its location on a computer network
UTF-8	Unicode Transformation Format – 8-bit, a variable-length character encoding used for electronic communication

1. Introduction

1.1. Background

Throughout the past decades we have observed a substantial shift of the traditional research applications in the field of natural sciences towards new modern techniques and automation. Species distribution modelling, ecological niche modelling, remote sensing and high-throughput sequencing technologies in genomics represent some of the breakthroughs that contribute to address high-priority issues such as the spread of infectious diseases, predicting the effects of global climate and land use change, effective conservation planning, global sustainability, world food and health security, as well as conserving ecosystem services. These issues are ultimately based on how the ecosystem really works, with complex processes and biological interactions that take place over a wide range of time scales and hierarchical levels of organisation (from genes to ecosystems) (Hardisty et al. 2013).

The recent technological advances in genomics such as DNA barcoding (Hebert and Gregory 2005) coordinated by the International Barcode of Life (iBOL), whole genome sequencing, the emergence of proteomics and metabolomics, new imaging methods (e.g. Computer Tomography or CT) (Hardisty et al. 2020), have produced significant quantities of data which along with chemical, morphological and geo-spatial information, bring the traditional “species” term to a higher level of knowledge. This assemblage of biodiversity data is often derived from specimen records held in Natural Science Collections (NSC) (Koureas and Addink 2017). NSCs represent a unique resource for scientific research in multiple ways. They are interconnected creating a worldwide interdisciplinary network that has enabled the emergence of modern uses of biodiversity data and an ever-increasing number of new users from various scientific fields.

The physical specimens as well as the various information bits derived from them, or from the context of collection (e.g. geographic location, elevation, habitats, collectors), are stored in digital form and provided by data repositories often in the context of Research Infrastructures that act on national, continental and global scales. Taxonomy as well as resolution of collector names have traditionally been used for the identification of physical specimens and biological taxa. While the inherent instability in taxonomic nomenclature is the nature of research activity, at the same time, it poses a great challenge when the need for linking specimens to highly scattered, derived information such as: images, DNA sequence data, trait measurements etc., emerges. The availability of linkages is essential for answering many scientific questions, for example, the construction of a phylogenetic tree highlighting the evolutionary relationships among taxa by utilising nucleotide sequences extracted from vouchered specimens.

The all-encompassing goal of the Biodiversity Community Integrated Knowledge Library (BiCIKL) project, funded by the European Commission, is to bring together infrastructures actively present in the biodiversity data landscape through the liberation of data from scholarly publications and bi-directional linking through persistent identifiers between literature, taxonomic, DNA sequence and occurrence data (Penev et al. 2021, Penev et al. 2022). Infrastructures participating in BiCIKL have been jointly working with the purpose of ensuring that their data will comply to the FAIR principles of Findable, Accessible, Interoperable and Reusable, with the ultimate goal of making these data of higher practical relevance when it comes to research and informing policy decisions, e.g. when tackling global environmental challenges. This has great implications for the end-users as they will be given

consistent access to enriched data and knowledge without having to search multiple databases, make an effort to link different data types or manually extract them from (a large number of) publications.

The bi-directional linking of several, different data types is a two-step process. Firstly, it requires the development and global adoption among all relevant stakeholders, of identifiers that are expected to be a) unique, b) persistent, c) always direct to a specific object and all the information linked to it, and d) actionable; they provide access to the identified object via mechanisms and services used by clients. These identifiers should ensure the creation of linkages to digital specimen objects, collections, taxonomic treatments, people, sequence data and taxa. Persistent Identifiers or PIDs offer fast access to data for various users, from researchers to policy-makers, while satisfying the FAIR Guiding Principles of being Findable, Accessible, Interoperable and Reusable[1]. The interaction with PIDs has been envisaged as being reliable and stable over long periods of time, regardless of continuous future technological advancements.

Secondly, on a technical level, the infrastructures hosting the interlinked biodiversity data are preoccupied with their generation, management and curation. An important part of this is the sharing of global biodiversity data. Application Programming Interfaces (APIs) are chosen as the means to present these data and when they adhere to best practices as they have been formulated by the biodiversity informatics landscape, they have the ability to address many of the possible challenges derived from data sharing such as standardised vocabularies, interoperability of heterogeneous data sources and data quality assessment (Norton, 2021). Towards this direction, APIs should be simple, user-friendly, pragmatic and designed to meet the needs of all user groups such as infrastructures, data providers, aggregators, developers etc. (Anderson et al, 2020, Norton, 2021). Effective communication with all stakeholders, easy to use technical solutions to the aforementioned challenges as well as working in teams are considered essential for the successful design and deployment and overall implementation of an API (Norton 2021).

1.2. Scope

This document outlines the best practices that are essential for each step of the interlinked biodiversity data life cycle, along with guidelines for the findability, re-usability and accessibility of infrastructures. It includes best practices with recommendations for findability of data and services, API provision, bi-directional linking, PIDs and recommendations on how to use the infrastructure services. Intended users are infrastructures, researchers, data providers and users of data aggregated in the infrastructures, and other end users like programmers and citizen scientists.

1.3. Structure of the document

The best practices are divided into six categories:

1) modalities of access,
2) building communities and trust,
3) technology and standards,
4) versioning of APIs and their data,

[1] https://www.go-fair.org/fair-principles/

5) bi-directional linking between infrastructures and
6) API design and naming conventions.

Each category lists best practices in blue, followed by recommendations for implementation in green. A second division of the same best practices into three user groups is presented in Appendix I for inclusion in the Biodiversity Knowledge Hub as 'information and guidelines' pages.

2. Findability, re-use and accessibility of infrastructures

2.1. Basis for the recommendations

The best practices with recommendations described in this document are based on:

- Recommendations for interoperability among infrastructures
- TDWG Biodiversity Services and Clients IG work
- Results from Technical RI Forum discussions regarding recommendations for APIs
- https://github.com/tdwg/apis/issues

2.2. Best practices with recommendations

Blue: best practice
Green: recommendation

1. Modalities of access

1.1. Primary scientific data needs to be provided as open as possible and only as closed as necessary for legal or sensitive data purposes.

1.1.1. It is recommended to provide metadata always under a public domain dedication (indicated as CC-Zero, or CC-0).

1.1.2. It is recommended to provide data under a public domain dedication or licensed under the Creative Commons that are Open Access compatible, e.g. CC-BY. NonCommercial (NC) or NoDerivatives (ND) licences are not recommended[2] for data intended for scholarly or scientific use, see: https://creativecommons.org/faq/.

1.1.3. It is recommended to provide the licence statement in a machine readable format. This allows search engines and software systems to be able to detect the CC licence. Machine readable HTML code for CC licences can be obtained from the CC licence chooser.

1.1.4. It is recommended to include a data quality assessment when data is provided.

[2] Horizon Europe allows CC-By NC and SA restrictions for 'long text' publications such as monographs:
https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/open-access-obligations-horizon-europe-what-are-cc-licences-2021-11-15_en

1.2. A Research Infrastructure providing data should ensure at a minimum a data discovery service plus CSV style data downloads, or RESTful endpoints to allow for programmatic access to the data. JSON is preferred over CSV because it can handle hierarchical information that the CSV format cannot.
  1.2.1. It is recommended to provide as many different modalities of access as possible, including access through packages for popular programming languages to work with data like Python or R. APIs can be used for different use cases requiring different kinds of APIs.
  1.2.2. It is recommended to provide APIs suitable for (future) machine-to-machine interaction, such as a DOIPv2 protocol implementation.
1.3. No person providing data should need to be contacted to obtain open access data except for cases like the need for very large amounts of data for which extraction through an API might not be efficient or appropriate.
1.4. Public APIs plus the data they serve should be fully documented and the documentation should be openly available and up to date.
  1.4.1. It is recommended that the API documentation (e.g. OpenAPI) covers common use cases and provides examples.
  1.4.2. It is recommended to provide machine-readable documentation, e.g. by using OpenAPI 3.x which can display the documentation both in a human readable (HTML) format and a machine readable (JSON) format.
  1.4.3. It is recommended to provide human-friendly descriptions and a beginners guide to the API(s).
  1.4.4. It is recommended to document the API versioning strategy and that versioning strategy should be precisely followed.
1.5. Public APIs served by a research infrastructure should be easy to find.
  1.5.1. It is recommended to provide multiple ways to discover public APIs such as listing them on the RI's website and registering them in dedicated service catalogues.
  1.5.2. At a minimum, the links to the API(s) and their documentation should be displayed on the RI's website for a straightforward discovery and access to the service.

2. Building communities and trust

2.1. APIs must be simple, easy to use, pragmatic, and designed with all major stakeholder groups in mind, including users, providers, aggregators, and architects.
  2.1.1. It is recommended to be as transparent as possible: every parameter in the request and response bodies should be defined and compromises should be thoughtful and documented.
2.2. Issues and requests for new API features should be easily reported and encouraged.
  2.2.1. It is recommended to have an open forum for issue reporting and discussion.
  2.2.2. It is recommended to provide development roadmaps openly.
  2.2.3. It is recommended to provide a mechanism for mass communication for developers to subscribe to notices about updates, downtimes, etc.
2.3. A mechanism for user support with clear response times should be provided.
  2.3.1. It is recommended to provide a free user support option.

2.3.2. It is recommended for public APIs to have a service status page providing information about e.g. historic uptime.

2.4. In case of write services, a sandbox or user acceptance test environment to allow users to contribute and test changes or to trial a service should be provided.

2.4.1. It is recommended for sandbox environments to indicate which part of the data is available with a clear policy on how to 'reset' the data.

2.4.2. It is recommended to make a clear distinction between data that is 'public' and data which (still) is under restricted access.

2.4.3. It is recommended to use a framework for service testing (such as JMeter).

2.4.4. In general, it is recommended to provide well tested services which build trust.

2.4.5. It is recommended to test performance of the service.

2.5. For data services (where sensible), a full dump of the (open) data served through the API at regular intervals (e.g. once a year) should be deposited in a trusted data repository

2.5.1. It is recommended to store data dumps with a DOI in a trustworthy data repository such as a CoreTrustSeal certified repository or Zenodo. To be trustworthy a data repository should follow the TRUST principles[3] for digital repositories.

2.6. Public APIs that require authorisation need to have a privacy policy describing how end-user data is processed.

2.6.1. It is recommended to include a fair use policy that describes when a service may be throttled to protect availability for other users.

2.6.2. It is recommended to protect personal data according to the Code of Conduct for Service Providers, a common standard for the research and higher education sector.

3. Technology and standards

3.1. Invest in standards compliance and work with organisations and communities to enhance existing standards or develop new ones.

3.1.1. It is recommended to provide data that adheres to the FAIR principles (having a PID, detailed metadata, data usage licence, etc).

3.1.2. For additional vocabularies that are in use with standards, it is recommended to have PIDs for the terms with their corresponding term descriptions.

3.1.3. For additional vocabularies that are in use with standards, it is recommended to have a clear process in place for proposing additional terms.

3.1.4. It is recommended to provide metrics that give credit to people (e.g. data providers) for work on standard compliance and development.

[3] Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

3.2. APIs providing biodiversity data need to use terms defined by the TDWG standards (e.g. DwC, ABCD, Audubon Core) if they are exact matches whenever possible[4].
  3.2.1. It is recommended to give preference to using DwC terms when similar alternatives in other standards exist.
  3.2.2. It is recommended to declare the namespaces with the terms. Terms from existing standards and vocabularies should be prefixed with their respective namespace abbreviations. Those prefixes should be defined and referenced accordingly. This allows for both humans and machines to easily identify linked terms.
3.3. Data needs to be provided in UTF-8 encoding if possible[5].
3.4. Data needs to be provided in a structured format.
  3.4.1. It is recommended to provide at least a JSON serialisation
  3.4.2. It is recommended to provide JSON as JSON-LD if it makes sense to do so, to conform to Linked Open Data. It is often not suitable for biodiversity datasets though[6]. However, JSON-LD can be used to format request and response metadata, just not the data itself.
  3.4.3. It is recommended that JSON responses are formatted following a published set of best practices such as those established by IIIF or OGC and that the design is consistent with it. For relational data it is recommended to use JSON:API, which specifies that endpoints are named as nouns rather than verbs. JSON:API is more a schema than a set of best practices though.
  3.4.4. It is recommended to serve data as 'flat' as possible, e.g. having at maximum two levels of nesting in JSON responses.
  3.4.5. It is recommended to support staged ('chunked') or queued (asynchronous) upload or download of very large files (where appropriate).
3.5. RESTful services need to properly use/recognize HTTP headers for requests and responses and return correct HTTP response codes accompanied with meaningful information in a human readable format. The API should return the status codes that cover all erroneous response types[7].
  3.5.1. Response format should be included using request headers rather than by expressing it in the URI.
  3.5.2. It is recommended that services provide content negotiation with at a minimum a serialisation in HTML for users (default) and in JSON for machines.
3.6. RESTful services requiring authentication need to provide access through HTTPS.
  3.6.1. It is recommended to always provide and require access through HTTPS rather than HTTP.
  3.6.2. It is recommended to use authentication only when really needed, such as for throttling or security.

[4] Reuse of terms is complicated. Terms that are not exact matches lead to unintentional consequences. See SKOS Mappings: https://www.w3.org/2004/02/skos/mapping/spec/2004-11-11.html
[5] Note that UTF-16 may be more efficient for e.g. Chinese characters where the 4 bytes needed in UTF-8 would take up twice as much space compared to UTF-16.
[6] For example, schema.org does not contain properties for taxonomic (e.g. phylum, class, family, subspecies) or trait (e.g., average mass, dietary preferences) data.
[7] http://docs.ogc.org/DRAFTS/19-072.html#http-status-codes

3.6.3. If authentication is required, it is recommended to provide it through OAuth2, e.g. through a token instead of the API getting the user's email address or password. There is a cost though: using OAuth2 can allow the third party provider access to the API activity.

3.7. RESTful service URIs need to indicate that they are part of an API either via a subdomain or a URL segment.

3.7.1. It is recommended that endpoints follow the naming conventions as specified in JSON:API and/or OGC API Specification.

3.8. For discoverability the API needs to be described such that the description can be indexed and found by search engines.

3.8.1. It is recommended to describe the API in Wikidata, using the Wikidata API Endpoint property: https://www.wikidata.org/wiki/Property:P6269.

3.8.2. It is recommended to create a service description using Schema.org for Web API: https://schema.org/WebAPI.

3.8.3. It is recommended to register a sitemap for landing pages describing the API.

4. Versioning of APIs and their data

4.1. API services should have an explicit version history with documentation about changes.

4.1.1. For REST-style APIs it is recommended to include the (major) version number in the URL path of the access point.

4.2. Data(sets) provided by API services need to have version information in its metadata, with a last modified timestamp or a date when the data was retrieved as minimum.

4.2.1. It is recommended to use an ISO 8601 date for last modified timestamps.

4.2.2. It is recommended to explicitly define the resources the API is built upon, indicate if they are updated & how.

4.3. Production versions of an API should be stable. APIs should rarely change as this may break existing implementations.

4.3.1. It is recommended not to change API endpoints. API endpoints should remain persistent, deprecation should be done through a versioning process where previous versions are preserved and the latest version are posted at a 'versioned' URL.

4.4. There should be a documented strategy for keeping older API versions online, e.g. with deprecation calendar/schedule.

5. Bi-directional linking between infrastructures

5.1. To enable bidirectional linking between infrastructures, resolvable PIDs need to be implemented for the data objects and provided through the APIs.

5.1.1. If direct linking cannot be supported between infrastructures, then it is recommended to use data brokers like wikidata to store links. Open linkage brokers provide a simple way to allow two-way links between infrastructures, without having to co-organize between many different organisations.

5.1.2. It is recommended to store created bi-directional links at both infrastructures between which the linkages are made.

5.1.3. It is recommended to provide provenance of the created linkages, who/what made the link, why and when.

5.2. APIs need to provide or accept as input identifiers traditionally held by other relevant organisations, including legacy identifiers where possible.

5.2.1. It is recommended to provide provenance information about established links in such a way that a data supplier can discover which of their data got enriched by linkages.

6. API design and naming conventions

6.1. APIs should provide predictive and consistent API behaviour. Some best practices that can be followed: https://iiif.io/api/annex/notes/design_principles/, https://www.w3.org/TR/ld-bp/,https://www.w3.org/TR/dwbp/,https://ogcapi.ogc.org/

6.1.1. It is recommended never to alter standard HTTP headers.

6.1.2. It is recommended to validate your response structures against a schema (where applicable).

6.1.3. It is recommended to use nouns instead of verbs in paths.

6.1.4. It is recommended to use easy to understand path elements in English; sometimes it may be beneficial to also use single vs plural e.g. /occurrences?query=... returning a list with occurrences versus /occurrence/123 returning one occurrence with id=123.

6.1.5. It is recommended that the technology used to produce endpoints to be hidden (eg /search vs /search.php).

6.1.6. It is recommended to make it RESTful, i.e., implement GET, PATCH, PUT, POST, DELETE, HEAD where relevant. for more information about RESTful API design see: https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

6.1.7. It is recommended to accommodate and use 301, 302, 303 redirects when possible and appropriate.

6.2. Request formats should be implemented in a non-ambiguous way.

6.2.1. It is recommended to enforce strict validation rules for request parameters and give hints if the validation fails (https://github.com/tdwg/apis/issues/32). Validation errors should be returned in both machine and human readable format with human readable instructions on how to rectify the error.

6.2.2. It is recommended to use a consistent request structure.

6.2.3. It is recommended to use only key value pairs for query parameters where possible.

6.3. The API provider should recommend how to atomise your data before you send individual requests.

6.4. The API provider should provide clear and identifiable responses.

6.4.1. It is recommended not to repeat response bodies. Each request should return a unique set of information in a response body. Two or more requests that return the same response body should be avoided.

6.4.2. It is recommended to include PIDs or GUIDs in response bodies where appropriate and within the proper context.

6.4.3. It is recommended to avoid deeply nested responses.

6.4.4. It is recommended to use HTTP response codes, and in addition show errors in human-readable text to provide some context (except for

context that provides sensitive information for e.g. hackers) and provide a verbose option where relevant.

6.4.5. It is recommended that, if a client requests a content type, to return that content type.

7. Some useful tips/tricks for developers

The following suggestions might considerably improve the API-based web services:

- For testing the API, it would be helpful to include an option during development for GET queries to include the query or request parameters in the response.
- You should include headers in responses in test mode.
- Some best practice documentation for API developers are to be considered as well:
    - https://en.wikipedia.org/wiki/List_of_HTTP_status_codes (response codes).
    - https://stackoverflow.blog/2020/03/02/best-practices-for-rest-api-design/ (RESTful design).

8. Suggestions to improve bi-directional linking for further exploration

While discussing best practices for findability, re-use and accessibility of infrastructures, several suggestions have been made for further investigation towards development of bi-directional linking between infrastructures. These are listed here for completeness:

- Options for using CETAF specimen identifiers when citing data through services like GBIF should be explored.
- Author guidelines on “How to cite specimens" (hyperlinked specimen IDs) when developed should be tested in pilot journals[8].
- The ways DiSSCo and INSDC/ENA could harvest and link back to literature citations of specimens and sequences, respectively, in their infrastructures, should be explored.
- Users may want to obtain the links and PIDs to the taxon name and all its synonyms when they search for it (especially in the case that different RIs use different taxonomic backbones such as when BOLD submits sequences to INSDC and they have to match taxon names).
- Material citations which contain sequences should be used to link these to the treatment name. Accession numbers mentioned within a material citation could provide a link between the specimen and the sequence taken. Links between specimens and sequences can also be managed in collection management software.
- It would be beneficial if a reference treatment could be added to each identification of a sequence.
- It should be considered to elaborate and publish annotated guidelines of how to publish material citations and tables including accession codes[9],[10].
- COL taxon names should be mapped together with the content in which they are described/detected (higher taxon categories), to promote usage by ecologists.
- It may be useful to develop a semantic, preferably event-based, model for linking, which should take into account legacy data from historical collections.
- It would be beneficial if standard identifiers for taxon names (or taxonomic concepts) such as those provided by COL, are used for specimens from different locations and if

[8] https://doi.org/10.3897/rio.8.e97374
[9] https://doi.org/10.3897/rio.8.e97374
[10] https://zookeys.pensoft.net/about#Linkeddatatableforprimarybiodiversitydata

taxon names are always linked to a verified taxon concept.
- Taxonomic treatments should also be cited along with taxon names since treatments connect a name (unique) or concept (unique) to specimens (multiple), establishing the unity between them.
- It would be useful to include type specimen information in taxon name aggregators such as COL and collections should provide the typified name of type specimens in their GBIF dataset.
- Online writing tools and publishing systems such as ARPHA should implement strict journal editorial policies and instructive guidelines to ensure that specimen PIDs are inserted by the authors.
- The visibility of the nomenclators (e.g. IPNI, or ZooBank) in COL should be raised and the addition of the links to the type specimens which are captured in the nomenclator databases should be explored.
- It would be good to harmonise the GBIF and COL name matching services and make these available for all datasets registered in the checklistbank.org.
- When a newly described species-level taxon is introduced it is highly recommended to link the record of a name in COL back to its primary publication source, treatment and holotype.
- It would be useful to have a mechanism to link a COL taxon name ID to the annual version make it available via API to automatically hyperlink a taxon name to its status in COL, e.g. Taxon name ID + year of its citation = disambiguated citation of a taxon concept from a particular annual version of COL.
- It would be useful if PIDs contain metadata about how the identifier should be cited.

3. Acknowledgements

We would like to thank everyone who reviewed and commented or otherwise contributed to this document. We would like to thank project partners who participated in the technical RI forum discussions towards this document, TDWG members who provided input and external experts, in particular Nicky Nicolson, Franck Michel and Sam Leeflang.

4. References

Agosti D, Benichou L, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8: e97374. https://doi.org/10.3897/rio.8.e97374

Anderson RP, Araújo MB, Guisan A, Lobo JM, Martínez-Meyer E, Peterson AT, Soberón JM (2020) Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. Frontiers of Biogeography 12 (3).https://doi.org/10.21425/F5FBG47839

Engelbrecht I, Steyn HM (2021) Does TDWG Need an API Design Guideline? Biodiversity Information Science and Standards 5: e75372. https://doi.org/10.3897/biss.5.75372 : Ian Engelbrecht, ian@nscf.org.za

Hardisty, A., Roberts, D. & The Biodiversity Informatics Community. A decadal view of biodiversity informatics: challenges and priorities. BMC Ecol 13, 16 (2013). https://doi.org/10.1186/1472-6785-13-16

Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6. https://doi.org/10.3897/rio.6.e54280

Hebert PD, Gregory TR. The promise of DNA barcoding for taxonomy. Syst Biol. 2005 Oct;54(5):852-9. https://doi.org/10.1080/10635150500354886. PMID: 16243770.

Koureas, Dimitrios & Addink, Wouter. (2017). Associating Occurrences with Genes, Phenotypes, and Environments through the Distributed System of Scientific Collections (DiSSCo). Proceedings of TDWG. https://doi.org/10.3897/tdwgproceedings.1.17010.

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

Meeus S, Addink W, Agosti D, Arvanitidis C, Balech B, Dillen M, Dimitrova M, González-Aranda JM, Holetschek J, Islam S, Jeppesen TS, Mietchen D, Nicolson N, Penev L, Robertson T, Ruch P, Trekels M, Groom Q (2022) Recommendations for interoperability among infrastructures. Research Ideas and Outcomes 8: e96180. https://doi.org/10.3897/rio.8.e96180

Michel F, Ettorre A, Faron C, Kaplan J, Gargominy O (2021) Biodiversity Knowledge Graphs: Time to move up a gear! Biodiversity Information Science and Standards 5: e73699. https://doi.org/10.3897/biss.5.73699 : Franck Michel, franck.michel@inria.fr

Norton B (2021) APIs: A Common Interface for the Global Biodiversity Informatics Community. Biodiversity Information Science and Standards 5: e75267. https://doi.org/10.3897/biss.5.75267 : Ben Norton (ben.norton@naturalsciences.org)

Patiiuk M (2021) A Case Study of Publishing Internal APIs to External Users. Biodiversity Information Science and Standards 5: e75386. https://doi.org/10.3897/biss.5.75386 : Max Patiiuk, max@specifysoftware.org

Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Barov B, Hobern D, Banki O, Addink W, Kõljalg U, Ruch P, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J (2021) Towards Interlinked FAIR Biodiversity Knowledge: The BiCIKL perspective. Biodiversity Information Science and Standards 5: e74233. https://doi.org/10.3897/biss.5.74233

Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin CS, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8: e81136. https://doi.org/10.3897/rio.8.e81136

Appendix I
Best Practices formatted for BKH

This appendix provides the best practices and recommendations ordered by user groups (Infrastructures, Data Providers, Users), in preparation for publication in the Biodiversity Knowledge Hub (BKH).

Infrastructures

All best practices and recommendations are relevant for Infrastructures.

Data providers

- Primary scientific data needs to be provided as open as possible and only as closed as necessary for legal or sensitive data purposes. ***(1.1 Modalities of access)***
    - It is recommended to provide metadata always under a public domain dedication (indicated as CC-Zero, or CC-0).
    - It is recommended to provide data under a public domain dedication or licensed under the Creative Commons that are Open Access compatible, e.g. CC-BY. NonCommercial (NC) or NoDerivatives (ND) licences are not recommended[11] for data intended for scholarly or scientific use, see: https://creativecommons.org/faq/.
    - It is recommended to provide the licence statement in a machine readable format. This allows search engines and software systems to be able to detect the CC licence. Machine readable HTML code for CC licences can be obtained from the CC licence chooser.
- Invest in standards compliance and work with organisations and communities to enhance existing standards or develop new ones. ***(3.1 Technology and standards)***
    - It is recommended to provide data that adheres to the FAIR principles (having a PID, detailed metadata, data usage licence, etc).
    - For additional vocabularies that are in use with standards, it is recommended to have PIDs for the terms with their corresponding term descriptions.
    - For additional vocabularies that are in use with standards, it is recommended to have a clear process in place for proposing additional terms.
- Data needs to be provided in UTF-8 encoding if possible. ***(3.3 Technology and standards)***
- Data needs to be provided in a structured format. ***(3.4 Technology and standards)***
- For discoverability the API needs to be described such that the description can be indexed and found by search engines. ***(3.8 Technology and standards)***
- To enable bidirectional linking between infrastructures, resolvable PIDs need to be implemented for the data objects and provided through the APIs. ***(5.1 Bidirectional linking between infrastructures)***

[11] Horizon Europe allows CC-By NC and SA restrictions for 'long text' publications such as monographs: https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/open-access-obligations-horizon-europe-what-are-cc-licences-2021-11-15_en

Users (Researchers, Developers, Citizen scientists)

- Primary scientific data needs to be provided as open as possible and only as closed as necessary for legal or sensitive data purposes. ***(1.1 Modalities of access)***
- A User providing data should ensure at a minimum a data discovery service plus CSV style data downloads, or RESTful endpoints to allow for programmatic access to the data. ***(1.2 Modalities of access)***
    - It is recommended to provide as many different modalities of access as possible, including access through packages for popular programming languages to work with data like Python or R. APIs can be used for different use cases requiring different kinds of APIs.
    - It is recommended to provide APIs suitable for (future) machine-to-machine interaction, such as a DOIPv2 protocol implementation.
- No person should need to be contacted to obtain open access data except for cases like the need for very large amounts of data for which extraction through an API might not be efficient or appropriate. ***(1.3 Modalities of access)***
- Public APIs plus the data they serve should be fully documented and the documentation should be openly available and up to date. ***(1.4 Modalities of access)***
    - It is recommended that the API documentation (e.g. OpenAPI) covers common use cases, on-the-fly request validation, meaningful error messages and provides examples.
    - It is recommended to provide machine-readable documentation, e.g. by using OpenAPI 3.x which can display the documentation both in a human readable (HTML) format and a machine readable (JSON) format.
    - It is recommended to provide human-friendly descriptions and a beginners guide to the API(s).
    - It is recommended to document the API versioning strategy and that versioning strategy should be precisely followed.
- Public APIs served by a User should be easy to find. ***(1.5 Modalities of access)***
- APIs must be simple, easy to use, pragmatic, and designed with all major stakeholder groups in mind, including users, providers, aggregators, and architects. ***(2.1 Building communities and trust)***
- Issues and requests for new API features should be easily reported and encouraged. ***(2.2 Building communities and trust)***
- A mechanism for user support with clear response times should be provided. ***(2.3 Building communities and trust)***
    - It is recommended to provide a free user support option.
- In case of write services, a sandbox or user acceptance test environment to allow users to contribute and test changes or to trial a service should be provided. ***(2.4 Building communities and trust)***
- For data services (where sensible), a full dump of the (open) data served through the API at regular intervals (e.g. once a year) should be deposited in a trusted data repository ***(2.5 Building communities and trust)***

Appendix II
Overview of infrastructure services with their access modes

Disclaimer: this overview was made to get an overview of the current services offered by the infrastructure in the BiCIKL project with their modes of access solely for the benefit of assessing the relevance of provided best practices and recommendations. It may be incomplete and not reflect the current situation as the infrastructure services are continuously evolving.

Table 1: Infrastructure services with their access modes

Name of infrastructure	Hosting institution	Service offered	Website	Identifiers and formats currently used	Modes of access
ARPHA-XML*	PENSOFT	XML-based manuscript	https://arphahub.com/	DOIs	search engine, filters
SIBiLS*	SIB	Automatic annotation pipelines and semantic search of full-text articles	https://candy.text-analytics.ch/SIBiLS/	JATS and BioC json	REST APIs
TreatmentBank*	PLAZI	Extraction, preparation and enhancement of data from literature	https://plazi.org/treatmentbank/	Treatment identifiers, DOIs, HTML, XML, TaxonX XML, RDF	full text search, search by taxonomic names/bibliographic data/material citations, PLAZI API
Meise Botanic Garden (MBG)*	MBG	specimens	https://www.plantentuinmeise.be/nl/home/	barcodes BR+13-digit number (e.g. https://www.botanicalcollections.be/specimen/BR5030011869353)	search engine, filters

Botanic Garden and Botanical Museum (BGBM)*	Freie Universität Berlin (FUB)	Herbarium specimens, botanical library	https://www.bgbm.org/en/biodiversity-informatics	barcodes B+number (e.g. http://herbarium.bgbm.org/object/B100326753)	search engine, filters
European Nucleotide Archive (ENA)*	EMBL-EBI	Nucleotide sequences	https://www.ebi.ac.uk/ena/browser/	XML, EMBL flat files, FASTA	ENA browser, free text search, detailed search with filters, accession search, API, large scale file download
Europe PMC*	EMBL-EBI	Life sciences literature	https://europepmc.org/	DOIs, PMIDs, PMCIDs	search engine, filters, RESTful APIs
OpenBiodiv*	PENSOFT	RDF-based biodiversity knowledge graph	http://openbiodiv.net	OpenBiodiv IDs (e.g. http://openbiodiv.net/365F75C1-1DC7-46E3-A07C-4793F8213B80) which correspond to triplets	API and SPARQL endpoint, search by taxa, treatments, sequences and other data elements
PlutoF*	UTARTU	Biodiversity data management and publishing platform (datasets that contain genomics, taxon names and OTUs, specimens, environmental samples, locality data, laboratory experiments, literature, interactions etc.)	https://plutof.ut.ee	DOIs, json	search engine for datasets, APIs

BE_VREs*	LifeWatch ERIC	Virtual Research Environments (VREs)	https://www.lifewatch.eu/catalogue-of-virtual-labs/		
LW_e-Infra*	LifeWatch ERIC	Data, services and VREs on Biodiversity and Ecosystem Research (BER)	https://lifewatch.eu		
BLR*	PLAZI	liberated and enhanced data from scholarly publications (access to PLAZI treatments, ZENODO datasets)	http://biolitrepo.org	Treatment identifiers, DOIs, XML, RDF	search engine with filters, API
Zenodo*	CERN	Liberated data and publications	https://zenodo.org/	ZENODO identifiers, json	search engine with filters, ZENODO rest API
DiSSCo*	Naturalis Biodiversity Center	Natural History collections	https://dissco.eu, https://sandbox.dissco.tech/	physical specimen identifiers, PIDs (e.g. Id: test/3bd56e68885ed1ae16ce), json	search engine, API (https://hdl.handle.net/api/handles/20.5000.1025/ZZX7-CEFZ)
CoL*	GBIF	Names and classification of species	https://www.catalogueoflife.org/	COL identifiers, json	Browse and Search tools, COL API

GBIF.org*	GBIF	Primary biodiversity data (occurrences, species, literature, datasets)	https://www.gbif.org/	GBIF identifiers, json	search engine with filters, GBIF API, rgbif
Biodiversity Heritage Library (BHL)**	Smithsonian Libraries and Archives	Biodiversity literature (including heritage-related publications)	https://www.biodiversitylibrary.org/	DOIs, BHL identifiers, PDF	search engine with filters
The Global Genome Biodiversity Network (GGBN)**		Samples, vouchers, taxa	https://www.ggbn.org/ggbn_portal/	catalog numbers, identifiers, GenBank numbers, Biorepository numbers, json	search engine with filters, GGBN API
The International Barcode of Life Consortium (iBOL)**		DNA barcode sequences	https://ibol.org/	Barcode Index Numbers (BINs), sample, sequence identifiers, GenBank accession numbers, XML, JSON, TSV, FASTA	search engine with filters, BOLD API
Metabarcoding Research and Visualization Environment (mBRAVE, iBOL)**		Projects based on high-throughput sequencing (HTS)	https://ibol.org/resources/informatics-platforms/		

Appendix III
API services compliance with described best practices

Disclaimer: This is an initial inventory of API services offered by the infrastructures in BiCiKL to get an indication of which best practices have already been implemented, solely to benefit the infrastructures in the project as a guide for improving their API services. It may not reflect the current state of the services though and may be incomplete. Also best practices and their recommendations may not be relevant for a certain service or there may be reasons to differ from a recommendation.

Table 2 : API services compliance with described best practices and recommendations

Name of infrastructure	modalities of access	building communities and trust	technology and standards	versioning of APIs and their data	bidirectional linking between infrastructures	API design and naming conventions	total
total number of best practices	5	6	8	4	2	4	29
ARPHA-XML	1.1, 1.2, 1.3	2.4	3.1				
SIBiLS	1.1, 1.2, 1.3, 1.4, 1.5	2.1, 2.3	3.1, 3.2, 3.4, 3.7, 3.8		5.2	6.1, 6.2, 6.3, 6.4	
TreatmentBank	1.1, 1.2, 1.3, 1.4, 1.5	2.1, 2.3	3.1, 3.2, 3.3, 3.4, 3.5, 3.7, 3.8	4.2, 4.3	5.1, 5.2	6.1, 6.2, 6.3, 6.4	
Meise Botanical Garden	1.1, 1.2	2.3	3.1				
Botanic Garden and Botanical Museum	1.1, 1.2, 1.3	2.3	3.1		5.1, 5.2		
European Nucleotide Archive	1.1, 1.2, 1.3, 1.4	2.1, 2.3	3.1, 3.2, 3.4, 3.5, 3.7	4.1, 4.2, 4.3	5.1, 5.2	6.1, 6.2, 6.3, 6.4	

Europe PMC	1.1, 1.2, 1.3, 1.4,	2.1, 2.2, 2.3, 2.5, 2.6	3.1, 3.2, 3.4, 3.5, 3.6, 3.7	4.1, 4.2, 4.3	5.1, 5.2	6.1, 6.2, 6.3, 6.4
OpenBiodiv	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3	3.1, 3.2		5.1, 5.2	
PlutoF	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3, 2.6	3.1, 3.2, 3.4, 3.5, 3.6, 3.7	4.2, 4.3	5.1, 5.2	6.1, 6.2
BE_VREs	-	-	-	-	-	-
LW_e-Infra	-	-	-	-	-	-
BLR	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3	3.1, 3.2, 3.4, 3.5, 3.7	4.1, 4.3	5.1, 5.2	6.1, 6.2, 6.3, 6.4
Zenodo	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3, 2.4, 2.6	3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7	4.1, 4.2, 4.3	5.1, 5.2	6.1, 6.2, 6.3, 6.4
DiSSCo	-	-	-	-	-	-
CoL	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3, 2.5	3.1, 3.2, 3.4, 3.5, 3.7	4.1, 4.2, 4.3, 4.4	5.1, 5.2	6.1, 6.2, 6.3, 6.4
GBIF.org	1.1, 1.2, 1.3, 1.4	2.1, 2.2, 2.3, 2.5, 2.6	3.1, 3.2, 3.3, 3.4, 3.5, 3.7	4.1, 4.2, 4.3, 4.4	5.1, 5.2	6.1, 6.2, 6.3, 6.4